Assessing Fit of the Lognormal Model for Response Times

Sandip Sinharay, Educational Testing Service
Peter van Rijn, ETS Global

An Updated Version of this document will appear in the Journal of Educational and Behavioral Statistics. The website for the journal is https://journals.sagepub.com/home/jebb

The citation for the article, as of 2/14/2020, is: Sinharay, S., & van Rijn (in press). Assessing fit of the lognormal model for response times. *Journal of Educational and Behavioral Statistics.*

Note: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170026. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education or of Educational Testing Service.

Assessing Fit of the Lognormal Model for Response Times

Sandip Sinharay, Educational Testing Service

Peter van Rijn, ETS Global

January 30, 2020

Note: Any opinions expressed in this publication are those of the author and not necessarily of Educational Testing Service.

Assessing Fit of the Lognormal Model for Response Times

Abstract

Response-time models are of increasing interest in educational and psychological testing. This paper focuses on the lognormal model for response times (van der Linden, 2006), which is one of the most popular response-time models. Several existing statistics for testing normality and the fit of factor-analysis models are repurposed for testing the fit of the lognormal model. A simulation study and two real data examples demonstrate the usefulness of the statistics. The Shapiro-Wilk test of normality (Shapiro & Wilk, 1965) and a Z-test for factor analysis models (Maydeu-Olivares, 2017) were the most powerful in assessing the misfit of the lognormal model.

Key words: χ^2 statistics, multivariate normal distribution, time intensity parameter.

Acknowledgements

The authors would like to thank the editors, Li Cai and Daniel McCaffrey, the associate editor, Rianne Janssen, and the three anonymous reviewers for several helpful comments that led to a significant improvement of the paper. The authors would also like to thank Jodi Casabianca-Marshall, John Donoghue, and Rebecca Zwick for several helpful comments, and James Wollack for generously sharing a data set that was used in the research that led to this paper. The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grant R305D170026. The opinions expressed are those of the authors and do not represent views of the Institute, the US Department of Education or the Educational Testing Service.

With the increasing popularity of computerized testing, which makes recording of response times straightforward, analysis of response times has become a rapidly expanding field of research. A common way to analyze response times is to include them in psychometric/statistical models that are referred to as *response-time models* (RTMs). The use of RTMs has been suggested to improve precision of examinee ability estimates (e.g. Bolsinova & Tijmstra, 2018; van der Linden, Klein Entink, & Fox, 2010), to detect test fraud (e.g., Qian, Staniewska, Reckase, & Woo, 2016; van der Linden & Guo, 2008; Sinharay & Johnson, 2019), to detect speededness (e.g., Schnipke & Scrams, 1997), to improve test construction (e.g. van der Linden, 2007), and to test substantive theories about cognitive processes (e.g., van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). Several RTMs have been suggested by, for example, Bolsinova and Tijmstra (2018), Klein Entink, Fox, and van der Linden (2009), Klein Entink, van der Linden, and Fox (2009), Maris (1993), Maris and van der Maas (2012), Rasch (1960), Schnipke and Scrams (1997), Thissen (1983), van der Linden (2006), van der Linden (2007), van der Maas et al. (2011), and Wang and Hanson (2005). Extensive reviews of RTMs include De Boeck and Jeon (2019), Kyllonen and Zu (2016), Lee and Chen (2011), Schnipke and Scrams (2002), van der Linden (2009), and van Rijn and Ali (2017).

The lognormal model for response times (LNMRT) is arguably one of the most popular RTMs. The model was first suggested by Thissen (1983), was further developed by van der Linden (2006), and has been considered, either to analyze only the response times, or to jointly analyze the response times and response accuracies, by several researchers including Bolsinova and Tijmstra (2018), Boughton, Smith, and Ren (2017), Glas and van der Linden (2010), Qian et al. (2016), Sinharay (2018), Sinharay and Johnson (2019), van der Linden (2007), van der Linden (2009), van der Linden and Glas (2010), and van der Linden and Guo (2008).

There is a lack of research on model-fit statistics for the LNMRT, Ranger and Kuhn (2014), Glas and van der Linden (2010) and van der Linden and Glas (2010) being among the few exceptions. This paper, in an attempt to fill that void, brings to bear several tools that have been used to assess fit of other statistical models to test item fit and the local

independence assumption for the LNMRT.

The next section includes a review of the LNMRT, existing approaches for estimation of the parameters of the model, and existing approaches for the assessment of fit of the model. The Methods section includes discussions of the model-fit statistics that we propose for assessing the fit of the LNMRT. The Simulations section includes an evaluation of the Type I error rate and the power of the statistics. The Real Data section includes applications of the statistics to two operational data sets. Discussions and conclusions are provided in the last section.

Reviews of the Lognormal Model, Fit Statistics, and Normality Tests

The Lognormal Response Time Model

The Model

Let us consider a test that includes J items. Let t_{ij} denote the response time, which is typically defined as the time an examinee spends on an item in a test, of examinee i on item j, where $i = 1, 2, \ldots, I,$, $j = 1, 2, \ldots, J$, and I is the number of examinees. Let us define

$$y_{ij} = \log(t_{ij}) \cdot$$

According to the LNMRT, y_{ij}'s, $j = 1, 2, \ldots, J$, are independent conditional on τ_i for any i and

$$y_{ij}|\tau_i \sim \mathcal{N}\left(\beta_j - \tau_i, \frac{1}{\alpha_j^2}\right), \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The parameter τ_i is the examinee's speed parameter; a larger value of the parameter results in smaller expected response time on all items for the examinee. The parameter β_j is the time-intensity parameter for item j; a larger value of the parameter results in larger expected response times for all examinees. The parameter α_j is the discrimination parameter for item j; a larger value of the parameter leads to more information on and hence smaller standard error of the examinee speed parameters. Given that the conditional mean of y_{ij} is $\beta_j - \tau_i$, one needs to impose a restriction on the model parameters to ensure

identifiability. This paper assumes that the population distribution $g(\tau_i)$ on τ_i is $\mathcal{N}(0, \sigma^2)$ for an unknown parameter σ^2, as in van der Linden and Guo (2008), which imposes the restriction that the population mean is 0; this restriction is similar to the restriction that the population mean of the examinee abilities is 0 in marginal maximum likelihood estimation in item response theory (e.g., Bock & Aitkin, 1981).

In applications of the LNMRT, researchers have analyzed only response times using the stand-alone LNMRT (e.g., Finger & Chee, 2009; Sinharay, 2018; van der Linden, 2006) or jointly analyzed both response times and response accuracies using the LNMRT and an item-response theory (IRT) model (e.g., Glas & van der Linden, 2010; van der Linden, 2007; van der Linden & Glas, 2010).

Estimation of the Item Parameters of the Model

A Markov chain Monte Carlo algorithm was suggested by van der Linden (2006) to estimate the parameters of the LNMRT. Glas and van der Linden (2010) suggested an approach to compute the maximum likelihood estimates (MLEs) of the item parameters when the LNMRT is used along with the three-parameter logistic model (3PLM) to jointly analyze both response times and response accuracy. Finger and Chee (2009) showed how one can use factor analysis to obtain the marginal maximum likelihood estimates (MMLE) of the item parameters of the stand-alone LNMRT and researchers such as Molenaar, Tuerlinckx, and van der Maas (2015) showed how one can use factor analysis to obtain the MMLEs of the item parameters of the joint model involving the LNMRT and an IRT model. Under the LNMRT, y_{ij} can be expressed as

$$y_{ij} = \beta_j - \tau_i + \epsilon_{ij}, \tag{2}$$

where ϵ_{ij}'s are independent of τ_i's, $E(\epsilon_{ij}) = 0$, and $\text{Var}(\epsilon_{ij}) = \frac{1}{\alpha_j^2}$. The above equation implies that

$$\boldsymbol{y}_i - \boldsymbol{\beta} = -\tau_i \mathbf{1} + \boldsymbol{\epsilon}_i, \tag{3}$$

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)'$, $\mathbf{1}$ is a $J \times 1$ vector of all 1's, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iJ})'$, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, and $\text{Var}(\boldsymbol{\epsilon}) = \boldsymbol{D}$, where the j-th diagonal element of the

$J \times J$ matrix $\boldsymbol{D}$ is equal to $\sigma^2 + \frac{1}{\alpha_j^2}$ and the off-diagonal elements of $\boldsymbol{D}$ are equal to σ^2. Equation 3 is like the equation of a factor-analysis model with one common factor τ_i where all the factor loadings are restricted to be equal to 1 (e.g., Joreskog, 1967).[1]

Therefore, we used the R package *lavaan* (Rosseel, 2012), which is used to perform factor analysis and structural equation modeling (SEM), to estimate the item parameters of the LNMRT.[2] The codes for using the *lavaan* package to compute the MMLEs for LNMRT are provided in Appendix A. Molenaar et al. (2015) noted that the LNMRT, when used as a component of a joint model, can be fitted using standard SEM software packages such as EQS, Lisrel, Mplus, and Mx.

The Need to Assess the Fit of the LNMRT as a Stand-alone Model

According to the Standard 4.10 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014), evidence of model fit should be documented when model-based methods are used. Therefore, it is important to assess the fit of the LNMRT. Given that the LNMRT was first presented as a stand-alone model by van der Linden (2006) and that the LNMRT has been used as a stand-alone model in various applications by, for example, Boughton et al. (2017), Marianti, Fox, Avetisyan, Veldkamp, and Tijmstra (2014), Qian et al. (2016), and Sinharay (2018), there is a need of more research on assessing the fit of the LNMRT as a stand-alone model. Also, a necessary condition of the joint model (consisting of the LNMRT and an IRT model) fitting both the response times and response accuracies is that the LNMRT fits the response times. Thus, if a simple test of fit of the LNMRT as a stand-alone model shows misfit, then one may not need to fit the joint model and instead can proceed with another RTM.

Given the two results that

- the effect of violation of the normality assumption is often negligible (e.g., Scheffe,

[1]The vector $\boldsymbol{\beta}$ is like the mean vector in the factor-analysis model.

[2]In limited simulations (the results of which are not reported here), the MMLEs of the item parameters produced by *lavaan* were found to be very accurate.

1959, p. 337),

- all models are wrong but some are useful (Box & Draper, 1987, p. 54),

one wonders whether the LNMRT is always useful, in terms of yielding accurate/valid inferences, because of its underlying normality assumption, or whether some types of misfit of the LNMRT have practical consequences and hence threaten the validity of the inferences from the LNMRT.

A substantial number of applications of the LNMRT involve the detection of test fraud. For example, Boughton et al. (2017), Fox and Marianti (2017), Marianti et al. (2014), Qian et al. (2016), Sinharay (2018), Sinharay and Johnson (2019), and van der Linden and Guo (2008) used the LNMRT to detect various types of test fraud. Specifically, person-fit analysis, which is one of the six types of statistical methods that are used in practice to detect test fraud according to Wollack and Schoenig (2018), was performed using the LNMRT in Marianti et al. (2014), Fox and Marianti (2017), and Sinharay (2018). Let us study the behavior of the Bayesian person-fit test statistic l^t of Marianti et al. (2014) under the misfit of the LNMRT. Klein Entink et al. (2009) showed that the LNMRT does not adequately fit data simulated from the Box-Cox normal model. A data set of response times of 5,000 examinees to 80 items was simulated from the Box-Cox normal model.[3] Then the LNMRT was fitted to the data set and the l^t statistic (Marianti et al., 2014) was computed for all examinees. At 5% level of significance, the l^t statistic[4] showed a statistically significant misfit for about 8% examinees whereas the misfit percentage would be 5% if the LNMRT were fitted to data from the LNMRT or the Box-Cox normal model were fitted to the data set. Thus, about 3% more examinees (or 150 examinees in the sample of 5,000) would be erroneously flagged as possible cheaters when the poor-fitting LNMRT is used instead of the better-fitting Box-Cox normal model. If data with even a

[3]The ν parameter, which indicates the skewness of the corresponding distribution, for the items was set to a mix of values 0.1, 0.2, and 0.3—Klein Entink et al. (2009) found the estimates of ν to be close to these values for some items for a real data set.

[4]Whose Type I error rate has been found to be close to the nominal level by Marianti et al. (2014) and Sinharay (2018).

larger extent of misfit of the LNMRT are simulated (by setting ν larger than 0.3), then the percent of examinees erroneously marked as possible cheaters would be even larger. Given the serious consequences of false alarms in the context of detection of test fraud (e.g., Skorupski & Wainer, 2017, p. 347), this data example shows that the LNMRT may be less useful than desired when it does not fit the data and its misfit may have practical consequences in some contexts and provides one more reason of the assessment of the fit of the LNMRT in all applications of the model to real data.

Existing Approaches for Assessment of Fit for the LNMRT

Schnipke and Scrams (1999) suggested the use of graphical plots and the root mean squared error between the observed and predicted cumulative distribution function of response times to assess the fit of the LNMRT. Several model-fit statistics have also been suggested in the context of applications of the LNMRT as a component of a joint model. These include the several Lagrange Multiplier statistics including one to assess item fit (Glas & van der Linden, 2010), the Lagrange Multiplier statistic for assessing conditional independence of the responses and response times (van der Linden & Glas, 2010), and the item-fit statistics of Ranger and Kuhn (2014).[5] Some of these methods can be adapted to applications of the LNMRT as a stand-alone model. The few model-fit tools that have been suggested for the LNMRT as a stand-alone model include the Lagrange Multiplier test for assessing conditional independence of the response times (van der Linden & Glas, 2010) and the Bayesian residuals based on the posterior predictive distribution of response times (van der Linden & Guo, 2008).

The item-fit statistic of Glas and van der Linden (2010) is designed to test the null hypothesis H_0 that the response times for item j follow the LNMRT given by Equation 1 versus the alternative hypothesis H_1 that the response times follow the distribution

$$y_{ij}|\tau_i \sim \mathcal{N}\left(\beta_j - \tau_i + \delta\omega(\boldsymbol{y}_i^{-j}), \frac{1}{\alpha_j^2}\right), i = 1, 2, \ldots, I, \tag{4}$$

[5]Posterior-predictive person-fit tests have been suggested by Marianti et al. (2014) and Fox and Marianti (2017), but this paper does not consider person fit.

where δ is an unknown parameter that can be estimated from the data and $\boldsymbol{y}_i^{-j}$ is the collection of log-response times of the i-th individual on all items except Item j, and

$$\omega(\boldsymbol{y}_i^{-j}) = \begin{cases} 1 \text{ if the sum of the elements of } \boldsymbol{y}_i^{-j} < r, \\ 0 \text{ if the sum of the elements of } \boldsymbol{y}_i^{-j} \geq r, \end{cases}$$

where r is an appropriate cut point that divides the examinees in two groups of roughly equal sizes. The alternative hypothesis essentially states that the mean of the response time is larger (compared to what is expected under LNMRT) for the slow examinees and smaller for the fast examinees or vice versa. Then the item-fit statistic of Glas and van der Linden (2010), henceforth denoted as the LM_j statistic, is obtained as the Lagrange Multiplier statistic for testing the null hypothesis that δ is equal to 0.[6] For large sample sizes, the distribution of the LM_j statistic can be approximated by a χ^2 distribution with one degree of freedom (df) when the LNMRT fits the data.

To compute the item-fit statistic of Ranger and Kuhn (2014) for item j, one first fits the RTM to the available data and then divides the examinees into G groups based on their response times on an item. Then, one counts the number of examinees who belong to these groups—this results in a total of G counts. Let us denote the collection of these counts for item j as o_{j1}, o_{j2}, ..., o_{jG}. One then computes e_{jg}'s, which are the expected value of the o_{jg}'s, under the assumption that the joint model fits the data. One finally computes the statistic

$$T_j = \sum_{g=1}^{G} (o_{jg} - e_{jg})^2$$

that quantifies the extent of model fit in item j. Ranger and Kuhn (2014) proved that the distribution of T_j is a multiple of a χ^2 distribution.

In the Bayesian approach of van der Linden and Guo (2008), one determines if the response time of an examinee-item combination is substantially different from what is expected under the model. van der Linden and Guo (2008) showed that in applications of the LNMRT as a stand-alone model, the posterior distribution of the predicted value of y_{ij}

[6] Glas and van der Linden (2010) described the test using a vector-valued δ, but used a scalar δ in their data examples—we describe the test using a scalar δ for simplicity.

conditional on $\boldsymbol{y}_i^{-j}$ is normal. Then, the standardized residual is computed as

$$e_{ij} = \frac{y_{ij} - E(y_{ij}|\boldsymbol{y}_i^{-j})}{\sqrt{\text{Var}(y_{ij}|\boldsymbol{y}_i^{-j})}}. \tag{5}$$

If the absolute value of e_{ij} is larger than an appropriate quantile of the standard normal distribution, the response time for the examinee on item i is concluded as aberrant. One can compute the e_{ij}'s for an item over all examinees and then conclude misfit for the item if the number of statistically significant standardized residuals for the item is larger than an appropriate cutoff (as in, for example, Boughton et al., 2017, p. 184). In our simulations (to be discussed later), this approach to assess item fit was found to have smaller power than the other item-fit statistics—so this approach is not considered henceforth.

van der Linden and Glas (2010) stated that under violation of local independence for item pairs $\{j, k\}$, the response times on these items for examinee i will follow a bivariate distribution given by

$$f(y_{ij}, y_{ik}) = \frac{\alpha_j \alpha_k}{2\pi\sqrt{1-\rho_{jk}^2}} \exp\left\{\frac{-1}{2(1-\rho_{jk}^2)}(\psi_{ij}^2 + \psi_{ik}^2 - 2\rho_{jk}\psi_{ij}\psi_{ik})\right\}, \tag{6}$$

where $\psi_{ij} = \alpha_j(y_{ij} - \beta_j + \tau_i)$. They defined a Lagrangian multiplier statistic to test for local independence of response times, which implies that ρ_{jk} is equal to 0, as

$$LM_{jk} = \frac{(\sum_i \hat{\psi}_{ij}\hat{\psi}_{ik})^2}{\sum_i \left(\hat{\psi}_{ij}^2 + \hat{\psi}_{ik}^2 - 1 - \frac{(\alpha_j\hat{\psi}_{ik} + \alpha_k\hat{\psi}_{ij})^2}{\sum_m \alpha_m^2}\right)}, \tag{7}$$

where $\hat{\psi}_{ij} = \alpha_j(y_{ij} - \beta_j + \hat{\tau}_i)$ and $\hat{\tau}_i$ is the MLE of τ_i and is given by

$$\hat{\tau}_i = \frac{\sum_j \alpha_j^2(\beta_j - y_{ij})}{\sum_j \alpha_j^2}. \tag{8}$$

The statistic LM_{jk} follows the χ^2 distribution with one degree of freedom when the local independence assumption holds.

Even though there exist several tools for testing the fit of the LNMRT, there seems to be a need for further research on assessing model fit in applications of the LNMRT as a stand-alone model. For example, there exists no comparison studies of the fit statistics

for the model. In addition, while there exist several χ^2-type statistics for assessing item fit for IRT models (e.g., Orlando & Thissen, 2000), there exist no χ^2-type statistics for assessing item fit for the LNMRT. This paper intends to fill this void by suggesting the use of several statistics to assess item fit for the LNMRT—these statistics have been used to assess the fit of other statistical models, but not to assess the fit of the LNMRT or any other response-time model.

Tests of Normality

Adefisoye, Golam Kibria, and George (2016) pointed to the existence of more than 40 tests of normality. These tests can be classified into the following three categories:

- tests based on empirical distribution function: Examples of such tests are the Kolmogorov-Smirnov test (e.g., Lilliefors, 1967), the Anderson-Darling test (Anderson & Darling, 1954), and the Lilliefors test (Lilliefors, 1967);
- tests based on moments: Examples of such tests are Jarque-Bera test (Jarque & Bera, 1980) and those based on skewness and kurtosis (e.g., Mardia, 1970);
- tests based on correlation: Examples of such tests are D'Agostino test (D'Agostino, 1971), Shapiro-Wilk test (Shapiro & Wilk, 1965), and Weisberg-Bingham test (Weisberg & Bingham, 1975).

In addition, as Thode (2002) noted, χ^2 type goodness-of-fit tests such as the Pearson's χ^2 test (Pearson, 1900) can also be used to test for normality.

There also exist several comparison studies of normality tests including Adefisoye et al. (2016), Gan and Koehler (1990), Razali and Wah (2011), Shapiro, Wilk, and Chen (1968), and Yazici and Yolacan (2007). Using data simulated from 45 different distributions and nine statistics, Shapiro et al. (1968) showed that the Shapiro-Wilk test statistic provided a general superior measure of non-normality. Seber (1984, p. 147-148) stated that among the tests of normality, the Shapiro-Wilk, Anderson-Darling, and D'Agostino test statistics are the most useful. Gan and Koehler (1990) compared the power of several normality tests and

found the Shapiro-Wilk test to be the best overall test for assessing normality. Yazici and Yolacan (2007) found three tests including the Jarque-Bera test to be the most powerful in a comparison of 12 tests of normality, but also found the Shapiro-Wilk test to be a superior omnibus indicator of normality. Razali and Wah (2011) compared the power of four tests—Shapiro-Wilk test, Anderson-Darling test, Lilliefors test, and Kolmogorov-Smirnov test—and found the Shapiro-Wilk test to be the most powerful. Adefisoye et al. (2016) found that the test based on kurtosis is the most powerful for observations from a symmetric distribution and the Shapiro-Wilk test is the most powerful for observations from an asymmetric distribution. The Nikulin-Rao-Robson test (Nikulin, 1973; Rao & Robson, 1974) statistic has also been found more powerful than tests of normality in detecting departures from normality (e.g., Voinov, Pya, & Alloyarova, 2009) and possesses several optimality properties (e.g., Singh, 1987; Voinov, Nikulin, & Balakrishnan, 2013, p. 37). In the simulation study later in this paper, the Shapiro-Wilk test, the Anderson-Darling test, and Nikulin-Rao-Robson test are used to test for item fit. Brief descriptions of these three tests are provided below.

Shapiro-Wilk Test

In a test for normality based on observations X_1, X_2, ..., X_I, the Shapiro-Wilk test statistic (Shapiro & Wilk, 1965) takes the form

$$W = \frac{\left(\sum_i a_i X_{(i)}\right)^2}{\sum_i \left(X_i - \bar{X}\right)^2}, \tag{9}$$

where $X_{(i)}, i = 1, 2, \ldots, I$, are a re-arrangement of the X_i's in an increasing order and the vector of weights a_i's, $\boldsymbol{a} = (a_1, a_2, \ldots, a_I)$'s, can be computed as

$$\boldsymbol{a} = \frac{\boldsymbol{m}'\boldsymbol{V}^{-1}}{\sqrt{\boldsymbol{m}'\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m}}},$$

where $\boldsymbol{m}$ and $\boldsymbol{V}$ respectively denote the mean vector and the covariance matrix of the standard normal order statistics. Tables of the critical values of the asymptotic null distribution of W can be found in, for example, Shapiro and Wilk (1965), and the statistic and its p-values can be computed using popular statistical software packages.

The p-values for the W statistic were computed in this paper by applying the R function shapiro.test (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/shapiro.test.html) on the logarithm of the response times for each item. A p-value smaller than the nominal level indicates an item misfit in the form of a departure from normality of the logarithm of the response times. Schnipke and Scrams (1999) used normal probability plots to assess fit of the LNMRT; however, the use of a normal probability plot involves a subjective judgment on misfit whereas the Shapiro-Wilk test statistic quantifies the misfit in a test statistic and a p-value—so the use of the Shapiro-Wilk test is more convenient, especially when these tests have to be done on a large scale.

Anderson-Darling Test

The Anderson-Darling test (Anderson & Darling, 1954) based on observations X_1, X_2, ..., X_I involves the use of the statistic

$$A^2 = -I - \frac{1}{I}\sum_i (2i-1)[\log \Phi(Y_i) + \log(1 - \Phi(Y_{I-i+1}))],$$

where $Y_i = (X_{(i)} - \bar{X})/s$, $\bar{X}$ is the mean of the X_i's and $s = \sqrt{\frac{1}{I-1}\sum_i (X_i - \bar{X})^2}$. Tables of the critical values of the asymptotic null distribution of A^2 can be found in, for example, D'Agostino (1986), and the statistic and the p-values for the statistic can be computed using popular statistical software packages. The p-values for A^2 were computed in this paper using the R function *ad.test* included in the R package *nortest* (Gross & Ligges, 2015) on the logarithm of the response times for each item. A p-value smaller than the nominal level indicates an item misfit in the form of a departure from normality of the logarithm of the response times.

Nikulin-Rao-Robson χ^2 Test

To apply χ^2 type goodness-of-fit tests to assess the normality for a sample of observations X_1, X_2, ..., X_I, the observations are divided into G groups based on the values of the X_i's so that the expected number of observations under the normal model is equal over the groups; such groups can be created by using group boundaries that are of the form

$\bar{X} + qs$, where $\bar{X} = \frac{1}{I}\sum_i X_i$ and q is a quantile of the standard normal distribution; for example, if 10 groups are used (that is, G=10), then the nine deciles of the standard normal distribution are used as the q's so that the group boundaries are $\bar{X} - 1.28s$, $\bar{X} - 0.84s$, $\ldots$, $\bar{X} + 1.28s$.[7] Then, O_g, the observed number of individuals belonging to Group g, is computed. The expected number of examinees in each group is $\frac{I}{G}$ because of the way the groups are constructed. The Pearson's χ^2 statistic represents the extent of departure of the expected and observed number of examinees falling in each group and is defined as

$$\chi^2 = \sum_g \frac{(O_g - \frac{I}{G})^2}{\frac{I}{G}} = \frac{G}{I}\sum_g O_g^2 - I\cdot$$

The Nikulin-Rao-Robson χ^2 statistic (Nikulin, 1973; Rao & Robson, 1974), which is also called the Rao-Robson χ^2 statistic, involves a correction term to the Pearson's χ^2 statistic and is defined as

$$\chi^2_N = \chi^2 + \frac{1}{I}\left(\sum_g \xi_g O_g\right)^2 + \frac{1}{I}\left(\sum_g \zeta_g O_g\right)^2, \tag{10}$$

where ξ_g's and ζ_g's are weights that depend on the normal density function and its derivative; their expressions can be found in Nikulin (1973). The χ^2_N statistic has an asymptotic χ^2_{G-1} null distribution and a large value of χ^2_N indicates a misfit of the normal model to the data.

Methods

In this section, several existing statistics for testing normality and the fit of factor analysis models are repurposed for testing item fit and the assumption of local independence of the LNMRT.

[7] $-\infty$ and ∞ are used as the lower boundary of the first group and the upper boundary of the 10th group, respectively.

Item Fit Analysis

Equation 1 implies that for a randomly chosen examinee, the marginal distribution of y_{ij} is given by

$$y_{ij} \sim \mathcal{N}\left(\beta_j, \sigma^2 + \frac{1}{\alpha_j^2}\right). \tag{11}$$

In addition, for an item, y_{ij}'s of the different examinees are independent. Therefore, tests of normality (e.g., Thode, 2002) can be used to assess item fit for the LNMRT. We used several tests of normality including the Shapiro-Wilk test, the tests for skewness and kurtosis, the Anderson-Darling test, the Jarque-Bera test and the Lilliefors test to assess item fit. Results for the Shapiro-Wilk test and the Anderson-Darling test are reported later. R codes for computing the p-values for these tests are provided in Appendix A. The other three tests had low power for the types of misfit considered later in our simulations—so the results of these statistics are not discussed henceforth.

We also used χ^2_N, the Nikulin-Rao-Robson χ^2 statistic (Nikulin, 1973; Rao & Robson, 1974), to test for item fit for the LNMRT. To compute the statistic, the number of groups, G, was set equal to 20, 30, and 50, respectively, for sample sizes of 500, 1,000, and 5,000 in our simulations; these values are in agreement with the recommendation of using $2I^{2/5}$ groups (e.g, Moore, 1986, p. 70) with χ^2-type tests; the use of other number of groups led to slightly smaller power of χ^2_N in a preliminary investigation. Moore (1986, p. 92) commented that "Among the chi-square statistics proposed and studied to date, the Rao-Robson statistic appears to have generally superior power and is therefore the statistic of choice for protection against general alternatives." Also, the χ^2_N statistic has been found to outperform the Pearson's χ^2 statistic in power comparisons (e.g., Rao & Robson, 1974; Voinov et al., 2009) and was more powerful than the Pearson's χ^2 statistic in our study—therefore results for the Pearson's χ^2 statistic are not provided. Also, given the optimality properties of the χ^2_N statistic (e.g., Singh, 1987), the statistic is expected to be more powerful than Ranger-Kuhn's χ_j statistic which, like χ^2_N, is computed from the differences between observed and expected frequencies, but, unlike χ^2_N, does not involve a division by the expected frequencies. R codes for computing the χ^2_N statistic for an item

are provided in Appendix A. There have not been too many comparison studies between normality tests such as the Shapiro-Wilks test and χ^2 tests for assessing normality of data. In addition, the advantage of using χ^2 tests to detect item fit of the LNMRT are that (a) they are in the same spirit as the χ^2-type statistics such as the Orlando-Thissen χ^2 statistics (Orlando & Thissen, 2000) for assessing item fit of IRT models, and (b) they may have more power than normality tests to detect some types of misfit.

Test for Local Independence

Several statistics for testing the local independence of the item scores, such as the Q_3 statistic (Yen, 1984), are based on the correlation between the scores on a pair of items. Borrowing the idea, a test statistic based on the correlation between the logarithm of observed response-times on a pair of items can be used to test the local independence of the response times.

Equation 1 and the population distribution $g(\tau_i) \equiv \mathcal{N}(0, \sigma^2)$ implies that

$$\begin{aligned} E(y_{ij}) &= E(E(y_{ij}|\tau_i)) = E(\beta_j - \tau_i) = \beta_j \\ \mathrm{Cov}(y_{ij}, y_{ik}) &= \mathrm{Cov}(E(y_{ij}|\tau_i), E(y_{ik}|\tau_i)) = \mathrm{Cov}(\beta_j - \tau_i, \beta_k - \tau_i) = \sigma^2 \quad (12) \\ \mathrm{Var}(y_{ij}) &= E(\mathrm{Var}(y_{ij}|\tau_i)) + \mathrm{Var}(E(y_{ij}|\tau_i)) = \frac{1}{\alpha_j^2} + \mathrm{Var}(\beta_j - \tau_i) = \frac{1}{\alpha_j^2} + \sigma^2. \quad (13) \end{aligned}$$

Let r_{jk} denote the observed correlation coefficient between the vectors of the examinees' log-response times on items j and k, where $j \neq k$. Given Equations 12 and 13, the corresponding population correlation coefficient under the LNMRT is

$$\rho_{jk} = \frac{\sigma^2}{\sqrt{\left(\sigma^2 + \frac{1}{\alpha_j^2}\right)\left(\sigma^2 + \frac{1}{\alpha_k^2}\right)}}. \quad (14)$$

The population correlation ρ_{jk} involves unknown parameters σ^2 and α_j's. Let $\hat{\rho}_{jk}$ denote the estimate of ρ_{jk}; one computes $\hat{\rho}_{jk}$ by replacing the parameters σ^2 and α_j's in Equation 14 by their estimates from a sample. For factor-analysis models, researchers such as Ogasawara (2001) and Maydeu-Olivares (2017) provided approaches to compute the estimated standard deviation $s(r_{jk} - \hat{\rho}_{jk})$ of the residual $r_{jk} - \hat{\rho}_{jk}$ and Maydeu-Olivares (2017) showed that

asymptotically,

$$Z_{LI} = \frac{r_{jk} - \hat{\rho}_{jk}}{s(r_{jk} - \hat{\rho}_{jk})} \sim \mathcal{N}(0, 1)$$

under the null hypothesis of the model fitting the data. Given the earlier discussion on how the LNMRT can be expressed as a factor-analysis model, the Z_{LI} statistic can be used to assess the local independence of the response-times for items j and k in applications of the LNMRT. A large absolute value of Z_{LI} would indicate the violation of local independence for the correponding item pair. The values of the Z_{LI} statistic were computed using the function *lavResiduals* in the R package lavaan (Rosseel, 2012). R codes for computing the Z_{LI} statistic are provided in Appendix A. Ogasawara (2001) and Maydeu-Olivares (2017) discussed other standardized residuals similar to Z_{LI}, but, in limited simulations, they were found to perform similar to Z_{LI}; therefore, results for only Z_{LI} among the residuals are provided in this paper. Molenaar et al. (2015) used modification indices to test for local independence of response-time models, but the Z_{LI} statistic provides a more rigorous and principled approach to test for the local independence for the LNMRT.[8] We also tried a version of Z_{LI} in which the denominator was the standard deviation of r_{jk}, rather than that of $r_{jk} - \hat{\rho}_{jk}$; while this version involves less computation compared to Z_{LI}, it had smaller power compared to Z_{LI} and is not considered henceforth.

Simulation Study

Three sets of simulations were performed to study the properties of and compare the performances of the following model-fit statistics: (a) Shapiro-Wilk statistic, (b) Anderson-Darling statistic, (c) Nikulin-Rao-Robson χ^2 statistic, (d) Ranger-Kuhn's T_j statistic (Ranger & Kuhn, 2014), (e) Lagrange Multiplier item-fit statistic LM_j (Glas & van der Linden, 2010), (f) Lagrange Multiplier local-independence statistic LM_{jk} (van der Linden & Glas, 2010), (g) Z_{LI} statistic (Maydeu-Olivares, 2017). The first set of simulations, which involved analysis of data generated under no model misfit, was

[8] While the modification index suggests whether a fixed parameter in a factor-analysis model can be freed (e.g., Satorra, 1989), it is not clear how the index can be used to test the hypothesis of local independence for an LNMRT.

intended to examine the Type I error rates of the aforementioned statistics. The second set of simulations, which involved analysis of data generated under some item misfit, was intended to examine the power of the item-fit statistics. The third set of simulations, which involved analysis of data generated under the violation of local independence, was intended to examine the power of the statistics for assessing local independence. Test lengths of 20, 40, and 60 items and sample sizes of 500, 1,000 and 5,000 were considered in each of the three sets of simulations.

Simulation of Data Under No Model Misfit

Data under no misfit were simulated under the LNMRT given by Equation 1. The true values of α_j's and β_j's were simulated from a $\mathcal{N}(1.87, 0.15^2)$ and a $\mathcal{N}(4, 0.45^2)$ distribution, respectively. The true values of τ_i's were simulated from a $\mathcal{N}(0, 0.3^2)$ distribution. These generating distributions were intended to make the summary of the simulated data resemble that of the real data described in van der Linden (2006). For example, with these generating distributions, the mean response times of the items were roughly between 25 and 150 seconds and the mean response times of the persons were between 20 and 170 seconds; these roughly match the corresponding quantities in Figures 1 and 2 of van der Linden (2006). A total of 100 data sets were simulated for each of the 9 combinations of test length and sample size—a new set of true values of the parameters was used to simulate each of these 100 data sets. For each simulated data set, the MLEs of the item parameters were computed using the lavaan package (Rosseel, 2012) and then these MLEs were used to compute the item-fit statistics and the statistics for assessing local independence. Statistical significance for each statistic was determined by comparing the values of the statistic to the corresponding theoretical percentile.

Simulation of Data Under Some Item Misfit

In this set of simulations, the generated data sets included a majority of examinees whose response times followed the LNMRT given by Equation 1 (the response times of these examinees were simulated in a manner similar to that for data simulated under no

model misfit) and a small fraction of aberrant/misfitting examinees whose response times did not follow the LNMRT. The percent of aberrant examinees in a data set was assumed to be 2, 5, or 10. Each generated data set included a majority of items with no misfit and a few items with some misfit. The number of items with misfit was assumed to be 2, 4, and 6, respectively, for test lengths of 20, 40, and 60, which means that misfit is assumed to be present for 10% of the items. The items with misfit and the aberrant examinees were randomly chosen for each simulated data set.

Three types of item misfit were considered including those arising from

- some examinees having preknowledge of the item
- the response times for the item being simulated from the Box-Cox normal model (Klein Entink et al., 2009)
- the response times for the item representing a positive shift for incorrect responses

To create the first type of misfit (item preknowledge), it was assumed that the response times of the aberrant examinees followed the LNMRT given by Equation 1 for the non-misfitting (or non-compromised) items, but were equal to 10, 20, or 30 seconds for the misfitting (or compromised) items (that constitute 10% of all items on the test). To create the second type of misfit (Box-Cox normal model), it was assumed that the response times of the aberrant examinees followed the LNMRT given by Equation 1 for the non-misfitting items, but were simulated from the Box-Cox normal model (Klein Entink et al., 2009) with ν-parameter 0.2, 0.5, or 0.8 for the misfitting items. To create the third type of misfit (positive shift for incorrect responses), the response times of the aberrant examinees for all the items were simulated from the LNMRT given by Equation 1, but a shift of 5, 10, or 15 seconds was added to their response times for the misfitting items only if their responses on the items were incorrect;[9] this type of misfit represents the scenario that those not knowing the answer to an item often spend more time on the item and eventually answer the item incorrectly—Ranger and Kuhn (2014) simulated this type of misfit in their simulation study.

[9]Item scores were generated for this case under the three-parameter logistic model.

For each simulation condition represented by a test length, a sample size, a percent of aberrant examinees, and a specific magnitude of misfit (represented by the time, ν-parameter, and shift for the three types of misfit), the following steps were iterated 100 times:

1. Simulate a data set with mostly non-aberrant examinees and some aberrant examinees;
2. Compute the MLEs of the item parameters for the data set;
3. Compute the item-fit statistics of the misfitting items using the MLEs computed above.

The power of each item-fit statistic for each simulation condition was computed as the percent of misfitting items that had a significant value of the statistic under that simulation condition.

Simulation of Data Under Violation of Local Independence

In this set of simulations, generated data sets included a majority of item-examinee combinations for which the response times followed the LNMRT given by Equation 1 (the response times of these examinees were simulated in a manner similar to that for data simulated under no model misfit) and some other item-examinee combinations for which local dependence was simulated in one of two ways.

To simulate the first type of local dependence, we assumed that 10, 20, or 40 percent of examinees suffer from speededness on one-fifth of the items at the end of the test and respond in 10, 20, or 30 seconds to those items. To simulate the second type of local dependence, we simulated response times for 10 item pairs using the bivariate distribution given by Equation 6 for 10, 20, or 40 percent of examinees. The correlation ρ_{jk} (that quantifies the extent of local dependence) in Equation 6 was assumed to be 0.05, 0.1, or 0.2.

For each simulation condition represented by a test length, a sample size, a percent of aberrant examinees, and a value of the number of items affected by speededness or a value of ρ_{jk}, the following steps were iterated 100 times:

1. Simulate a data set that involves violation of local independence;

2. Compute the MLEs of the item parameters for the data set;

3. Compute the LM_{jk} and Z_{LI} statistics for all the item-pairs for which the local independence assumption was violated.

Results for Data Simulated Under No Model Misfit

Except for the LM_j and LM_{LI} statistics, all the other statistics have satisfactory Type I error rates (Appendix B includes a table showing these rates)—the rates are close to and often smaller than the nominal level in all simulation conditions.

Results for Data Simulated Under Some Item Misfit

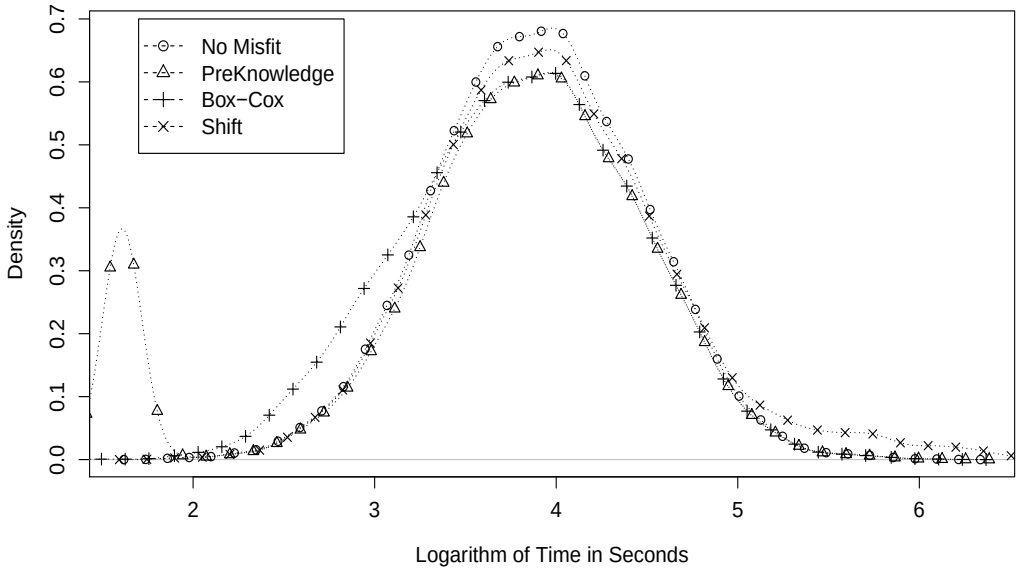

Figure 1: Density plots of logarithm of response times for four items.

Figure 1 shows the density plots of the logarithm of response times for four items from the simulation cases involving 5,000 examinees and 80 items—the LNMRT fits the data for Item 1 (circles on a dotted line) so that the logarithms of the response times follow the normal distribution for that item and does not fit the other items. Item 2 (triangles on a dotted line) represents an item on which 10% examinees had preknowledge and answered

the item in 5 seconds. The response times for Item 3 (plus symbols on a dotted line) were simulated from the Box-Cox normal model (Klein Entink et al., 2009). The response times for Item 4 (multiplication symbols on a dotted line) represents a shift of 5 seconds for the wrong responses. The figure shows that the distribution of the response times for Item 2 is bimodal. The figure also shows that compared to the item with no misfit, the distribution of the logarithm of response times for Item 3 is shifted slightly towards the left and that for Item 4 is shifted slightly towards the right. In addition, the distributions for Items 2-4 have lower peak compared to that of Item 1. The value of the Nikulin-Rao-Robson statistic is approximately 51, 2805, 77, and 534 for the items, the critical value at 5% level being 66.3 (that is the 95th percentile of a χ^2 distribution with 49 degrees of freedom).

Other factors remaining the same, the power of each statistic was very similar over different test lengths. Figures 2 to 4 respectively show the average power (averaging over the different test lengths) for detecting the three types of item misfit for different values of sample size (I), percent of aberrant examinees, and the extent of misfit (denoted by the time-taken-to-answer-the-compromised-items, ν-parameter, or the shift for the three types of misfit) of four of the item-fit statistics. The values of power at 5% level are reported in these figures; the conclusions from power at 1% level (not reported here) are very similar. The three rows of the figures correspond to sample sizes 500, 1,000, and 5,000, respectively. The three panels in each row show the average values of power of the statistics for various extent of misfit. In each panel, the percent of aberrant examinees is shown along the X-axis and the power of each statistic is shown along the Y-axis. The power for the Shapiro-Wilk (SW) statistic, Anderson-Darling (AD) statistic, Nikulin-Rao-Robson (NRR) statistic, and the Ranger-Kuhn's (RK) T_j statistic are shown using hollow circles, hollow triangles, plus signs, and multiplication signs, respectively, joined by a dotted line.

The power of the Lagrange multiplier item fit statistic (LM_j) is much smaller compared to the other statistics—hence this statistic is not included in Figures 2 to 4. The low power may be due to the fact that the misfit created in this paper is not the type of misfit that the statistic is ideal to detect. In limited simulations, the LM_j statistic was found to have larger power when item misfit was created by simulating data using Equation 4 instead of

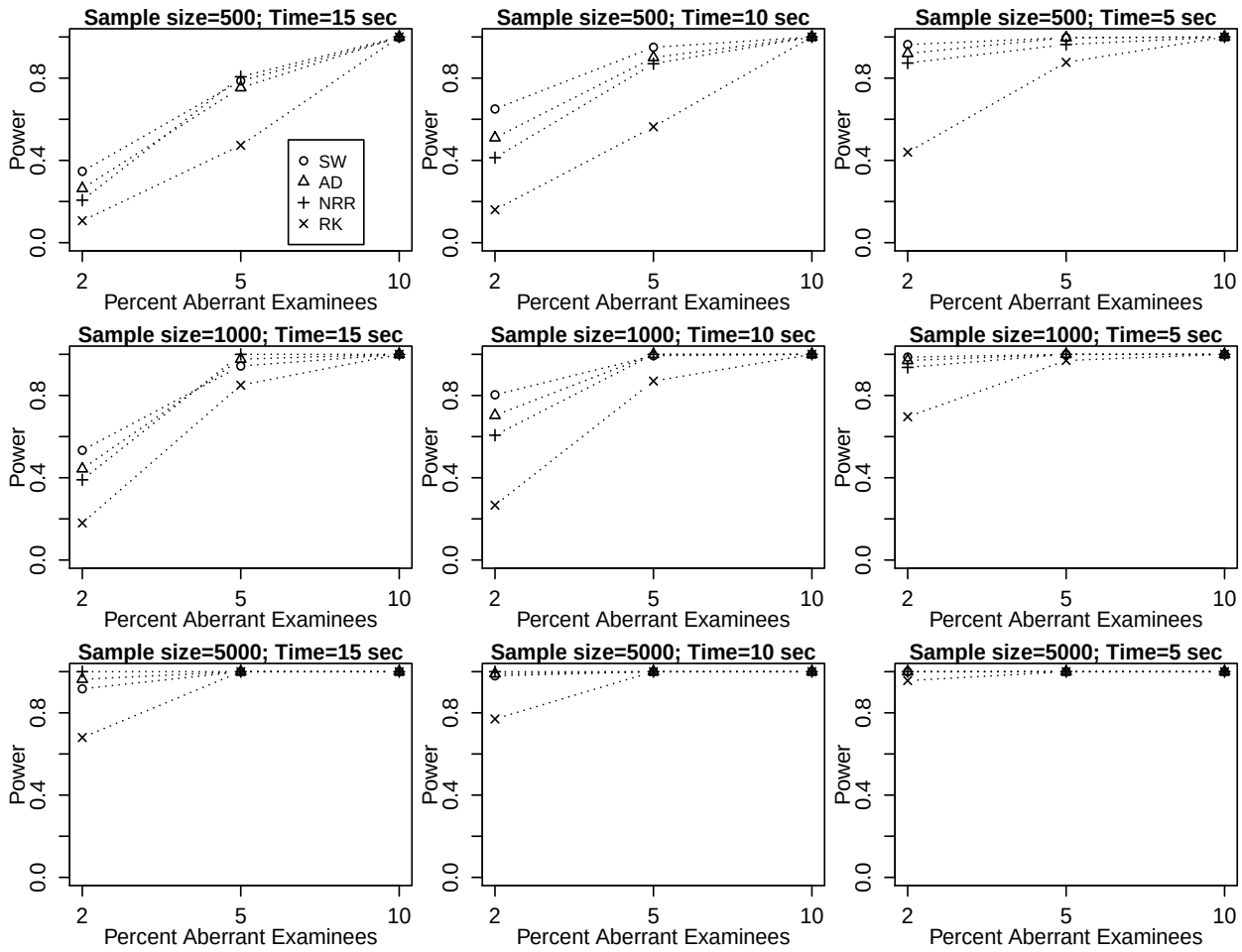

Figure 2: Average power across test lengths to detect the first type of item misfit (Preknowledge) of the item-fit statistics.

Equation 1.[10]

Figures 2-4 show that:

- Power increases with an increase in sample size, which is a favorable result for the item-fit statistics (e.g, Rao, 1973, p. 464).
- Power becomes larger as the percent of aberrant examinees increases.
- Power mostly becomes larger as the extent of misfit (denoted by time or the ν-

[10]However, from a study of the estimated residuals given by Equation 5 for our real data sets to be discussed later, it was not clear that misfit created by using Equation 4 is prevalent in practice—so this type of misfit is not considered in our simulation study.

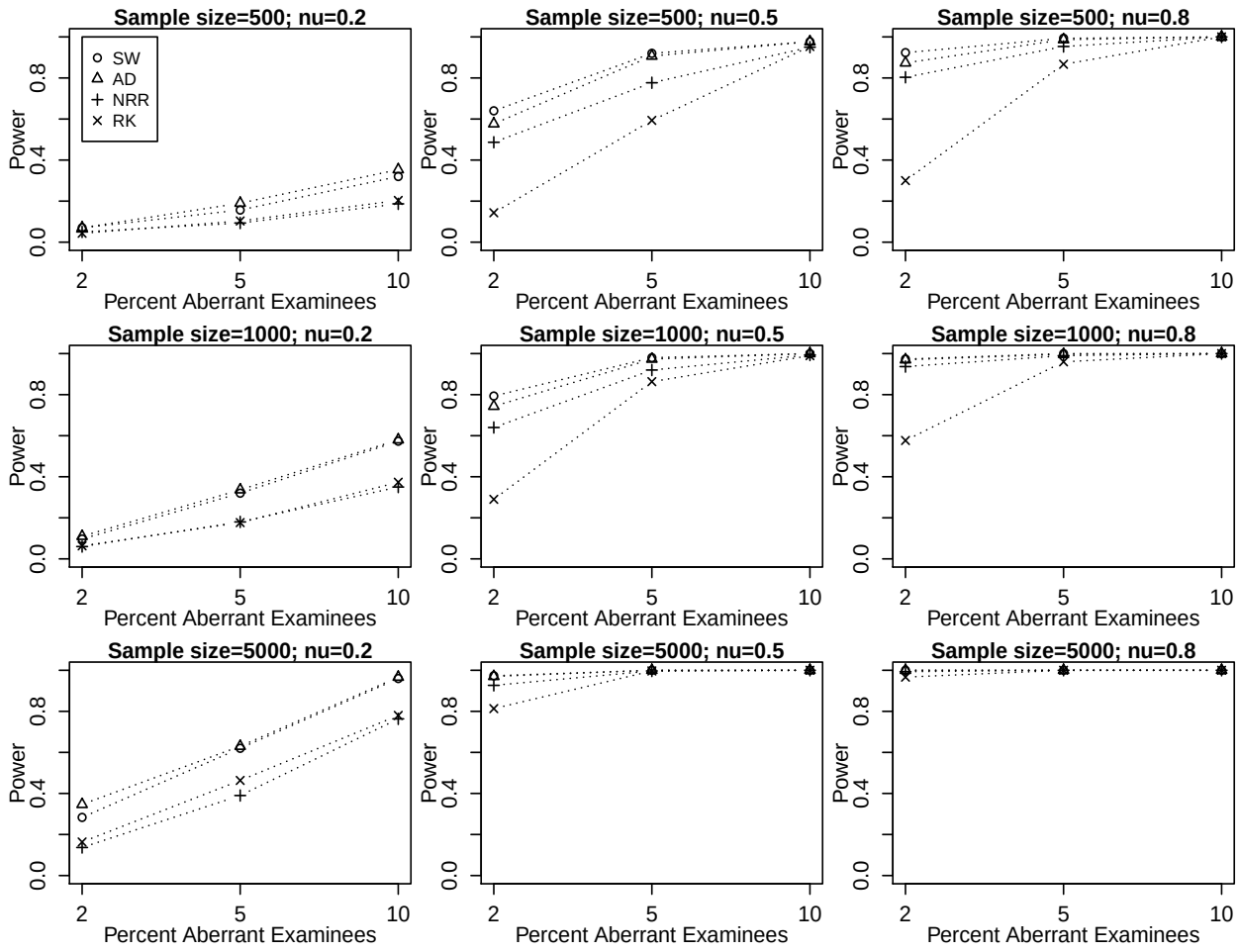

Figure 3: Average power across test lengths to detect the second type of item misfit (Box-Cox normal distribution) of the item-fit statistics.

parameter or the shift) increases.

- The power of all the statistics is very close to 1 in the bottom right panel (that is, for large samples and large extent of item misfit) in each figure.

- No item-fit statistic uniformly has the largest power in all simulation cases, but the Shapiro-Wilk statistic comes close to achieving this distinction. The statistic consistently has the largest power for small sample sizes and small percent-aberrant examinees. The large power of the Shapiro-Wilk statistic is in agreement with the superior performance of the Shapiro-Wilk statistic in several comparisons of normality tests (reviewed earlier).

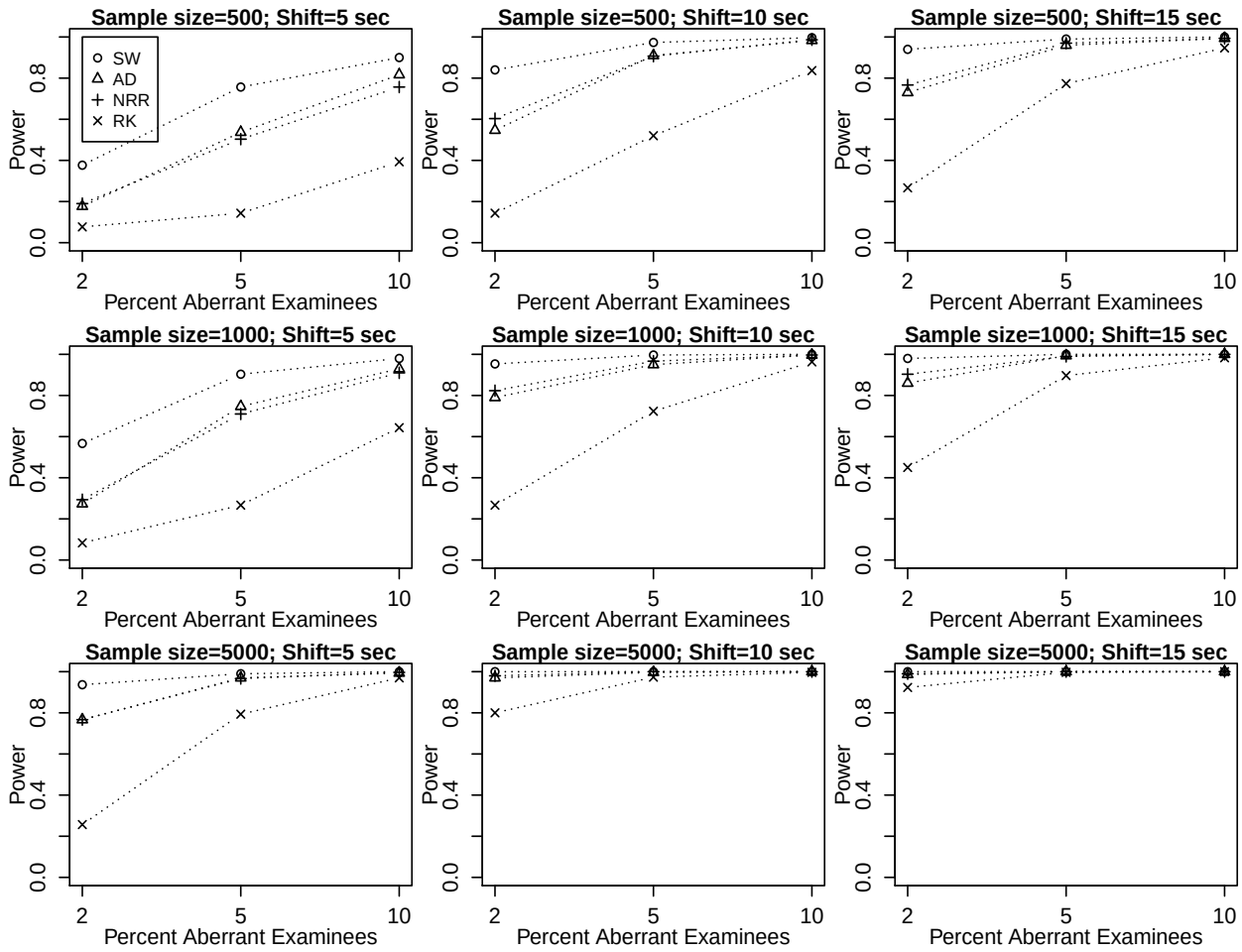

Figure 4: Average power across test lengths to detect the third type of item misfit (a shift in time for incorrect answers) of the item-fit statistics.

- The Nikulin-Rao-Robson χ^2 statistic has the largest power in a few cases (for example, in the bottom left and middle panels of Figure 2), but is less powerful than the Shapiro-Wilk and Anderson-Darling statistics in general.

- The Ranger-Kuhn statistic (T_j) has the smallest power overall.

Results for Data Simulated Under Violation of Local Independence

Figure 5 shows the average values of power (averaging over the test lengths) of LM_{jk} and Z_{LI} to detect violation of local independence of the two types in a similar manner as in Figure 2 except that the three panels in each row of the figure correspond to the three values of the response times under speededness (top row) or the three values (0.05, 0.1, and

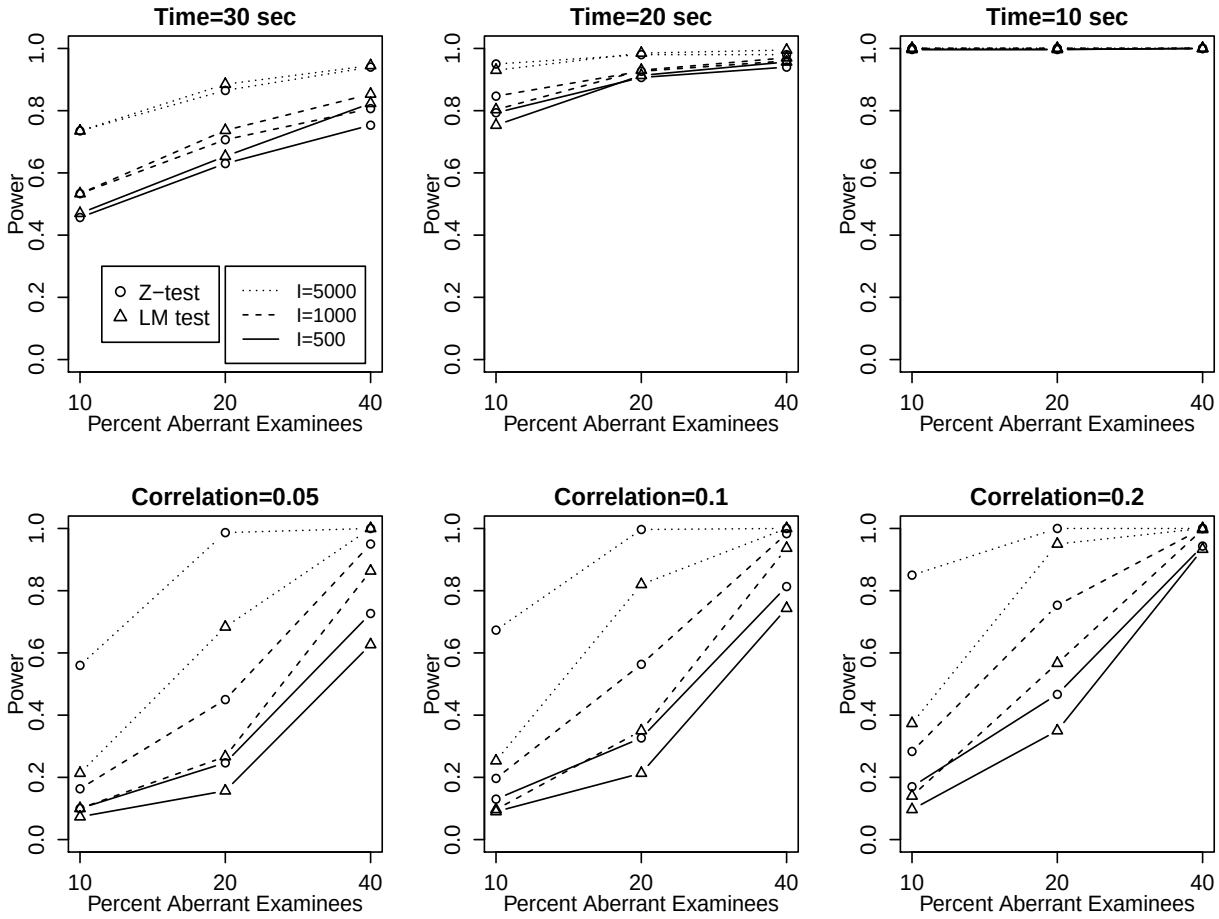

Figure 5: Power to Detect Violation of Local Independence.

0.2) of ρ_{jk} (bottom row) and each panel shows results for all sample sizes, using dotted lines (sample size of 5,000), dashed lines (1,000) and solid lines (500). Figure 5 shows that:

- The power of both statistics increases as the sample size or the percent of aberrant examinees increases.
- For the sample size of 5,000, the power of Z_{LI} is extremely high if the percent of aberrant examinees is 20 or larger.
- The power increases with a decrease in time (of responding under speededness) or an increase in ρ_{jk}.
- The Z_{LI} statistic is much more powerful than LM_{jk} in all simulation cases.

Real Data Examples

Example 1: A Mathematics Test

Let us consider a real data set that consists of responses and response times of 1,079 American test takers in grade 8 on 40 mathematics items. The data were analyzed by van Rijn and Ali (2017) and were collected as part of a larger study. Thirty-two items are multiple-choice and eight are numeric entry. The items focus on basic topics in number, measurement, geometry, data analysis and algebra, and are dichotomously scored. The items were assembled in four different forms using blocks of ten items, with different orders of the blocks to counterbalance order effects. The time limit was 90 minutes.

The test was administered under low-stakes conditions—so we computed the response time effort (RTE) measure of Wise and Kong (2005) for the data set to examine if the examinees suffered from a lack of motivation. The RTE for an examinee is the proportion of items for which the response time of the examinee is above a cutoff. With a cutoff of 5 seconds, a value of 0.8 of the RTE means that an examinee took more than 5 seconds on 80% of the items (so lower values of RTE mean less effort). With a cutoff of 5, 10 and 15 seconds, respectively, 0, 10 and 64 out of the 1079 examinees had an RTE less than .80. With a cutoff value of 10 seconds, the lowest RTE value found for the data set is .45, meaning that the corresponding examinee spent 10 seconds or less on 55% of the items. The number of examinees answering an item in less than 10 seconds range between 10 and 151 for the items. These numbers indicate that overall, there is not enough evidence that many examinees suffered from a lack of motivation. Figure 6 shows the total time in minutes versus the raw score on the test—it shows almost no correlation (correlation coefficient=-0.05) between the total time and raw score.

The MLEs of α_j's were between 1.18 and 2.05 and those of β_j's were between 2.91 and 4.96, respectively. Values of both the Shapiro-Wilk test statistic and the Nikulin-Rao-Robson χ^2 statistic were computed for all the items in the data set. The number of groups used was 20 for the latter test. At 5% level, all the items were found to have statistically significant values of both the statistics. At 1% level, 97.5% and 100% of

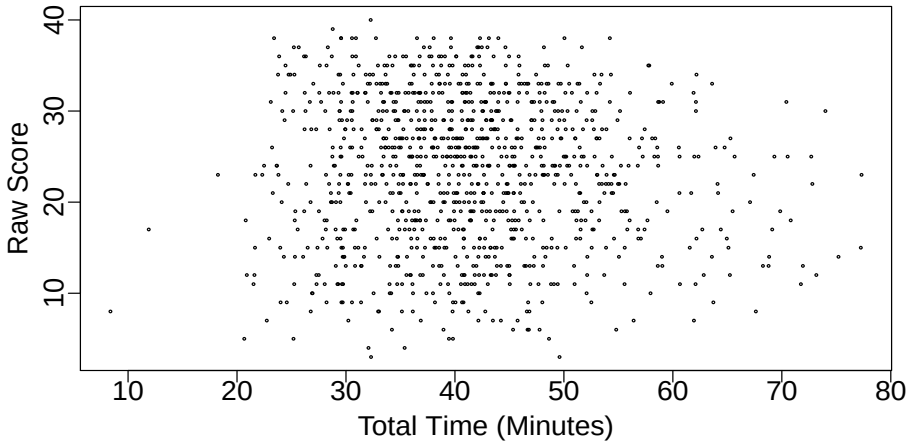

Figure 6: Plot of total time versus raw score for the Mathematics data.

the items were found to have statistically significant values of the Shapiro-Wilk statistic and the Nikulin-Rao-Robson statistic, respectively.

Figure 7 shows the normal probability plots of the log-response times for nine items (randomly chosen) among the 40 items. In each panel, a diagonal line is also shown for convenience—a curve close to the diagonal line would indicate a good fit of the LNMRT to the data. The figure shows that while the fit is not too bad in the right side of the panels, the curve drops well below the diagonal line towards the left of several panels. One cause of this drop is quick responding by several examinees. This is demonstrated in Figure 8. The left panel of the figure shows the standardized residuals e_{ij}'s (van der Linden & Guo, 2008) versus the estimated τ_i's for all the examinees for Item 8. Horizontal lines are drawn at 2 and -2—residuals outside these lines are statistically significant at 5% significance level. The right panel shows the actual response times versus the estimated τ_i's for all the examinees for the item—a logarithmic scale is used for the vertical axis because of the presence of several large response times. Horizontal lines are drawn at 10 and 20 seconds. The left panel shows many significant residuals (for about 6% of examinees)—most of these are negative, indicating that several examinees responded to the item much sooner than what can be expected under the LNMRT. The right panel shows that a substantial number

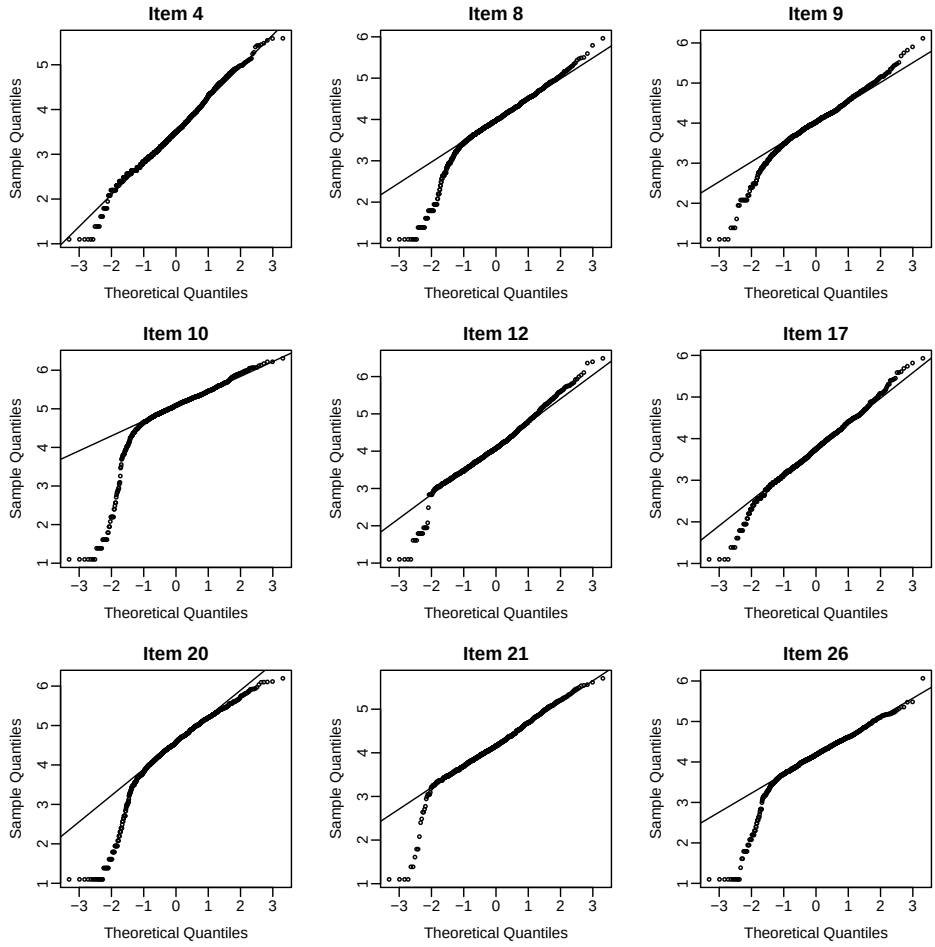

Figure 7: Normal probability plots for Nine items from the Mathematics data.

of examinees answered the item within 5-15 seconds, whereas the average response time for the item was 61 seconds, and justifies the way the first type of item misfit was created in our simulation study discussed earlier.

Figure 9 provides a deeper look at the relationship between the misfit and quick responding. The left panel shows a plot of the average per-item time (in seconds) of the examinees (X-axis) versus the values of a person-fit statistic using response times (Sinharay, 2018) whose larger values indicate more misfit[11] (Y-axis). A horizontal dashed line is provided at the critical value of the person-fit statistic at 5% significance level—values of the statistic above this line are significant. The correlation between the two plotted quantities

[11]A misfit for a person indicates that the LNMRT does not adequately fit the response times of that person.

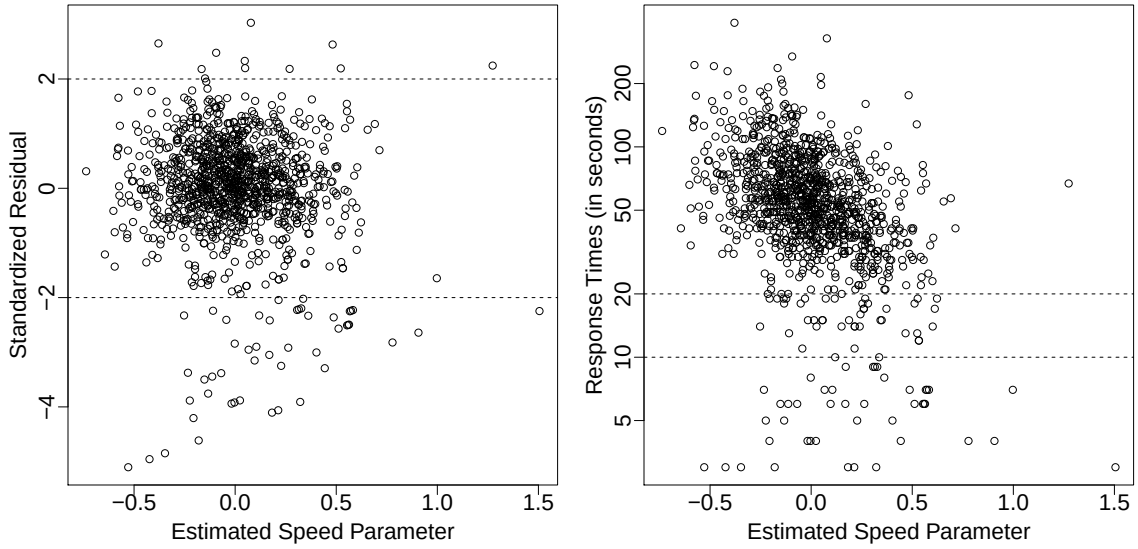

Figure 8: The residuals and response times versus the estimated speed parameters for Item 8.

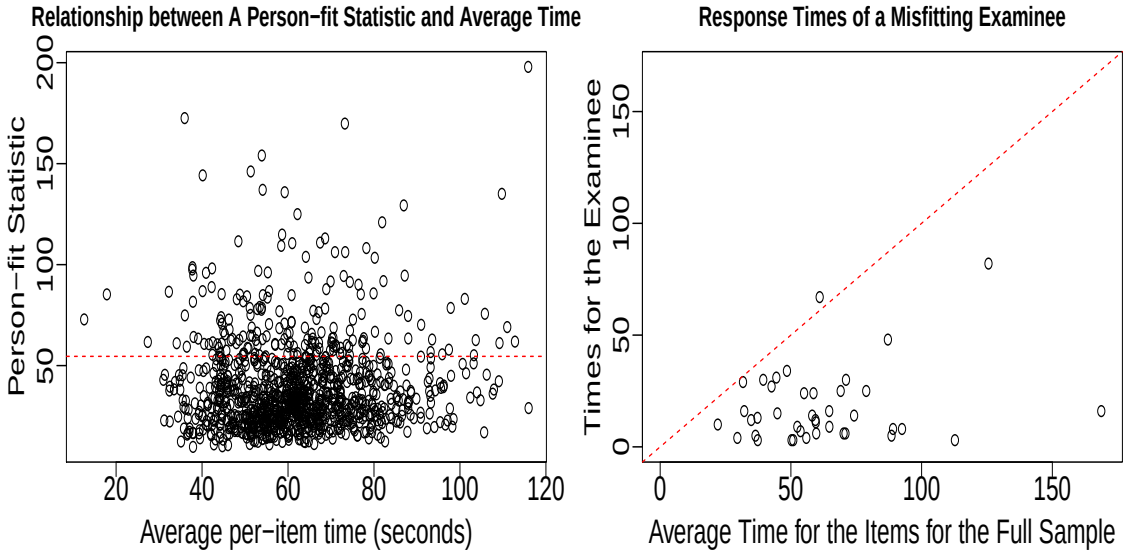

Figure 9: The relationship between misfit of the LNMRT and quick responding.

is -0.25, which, together with the plot, indicates that more misfit is associated with quicker responding. Especially, the three quickest examinees (who appear on the extreme left of the top panel) all have significant values of the person-fit statistic.[12] The right panel shows a plot of the average per-person time (in seconds) on the items (X-axis) for all examinees

[12]In addition, their raw scores are 8, 18, and 23, respectively—so it is not clear that they answered quickly because they were very strong in mathematics.

in the sample versus the response times (in seconds) on the items (Y-axis) for one of the examinees who appear on the extreme left of the top panel; a diagonal (dashed) line is added to the plot; the panel shows that the examinee answered all the items except one faster than the other examinees on average. Thus, Figure 9 shows that a part of the severe extent of item misfit in the data can be attributed to some examinees who responded quickly. But several points above the horizontal line and towards the right of the left panel of Figure 9 show that some examinees took longer than average and yet had significant values of the person-fit statistic—so quick responding is not the only source of the misfit of the LNMRT to the data.

In addition, the $S - \chi^2$ item-fit statistic based on item scores and suggested by Orlando and Thissen (2000) was computed for all the items—the statistic was significant for 7 items (or about 18% of the items) at 5% significance level. The correlation between $S - \chi^2$ and the Nikulin-Rao-Robson χ^2 statistic was 0.25—so there seems to be a small positive association between item fit based on item scores and item fit based on response times.

The value of the Z_{LI} statistic was significant at 5% level for 37.3% item-pairs. Figure 10 shows the values of the Z_{LI} statistic for the item-pairs. The figure was created using the R package *corrplot* (Wei & Simko, 2017). The item numbers are shown at the left and the top of the figure. A larger black or white square for a pair of items indicates a Z_{LI} for that pair that is large in absolute value. Black and white squares indicate statistically significant (at 5% significance level) and positive and negative values, respectively. If Z_{LI} for an item pair is statistically not significant at 5% level, then no square is drawn for that pair (and the background for that item pair remains gray). The several large black squares close to the diagonal indicate that the response-times for the item-pairs within each block (of 10 items) are more correlated than what is expected under the LNMRT. Especially, there are groups of black squares in the top left corner and the bottom right corner of the figure. There are several large white squares indicating, for example, that the response-times for an item-pair with one item among items 21-28 and another among items 11-20 are less correlated than what is expected under the LNMRT.

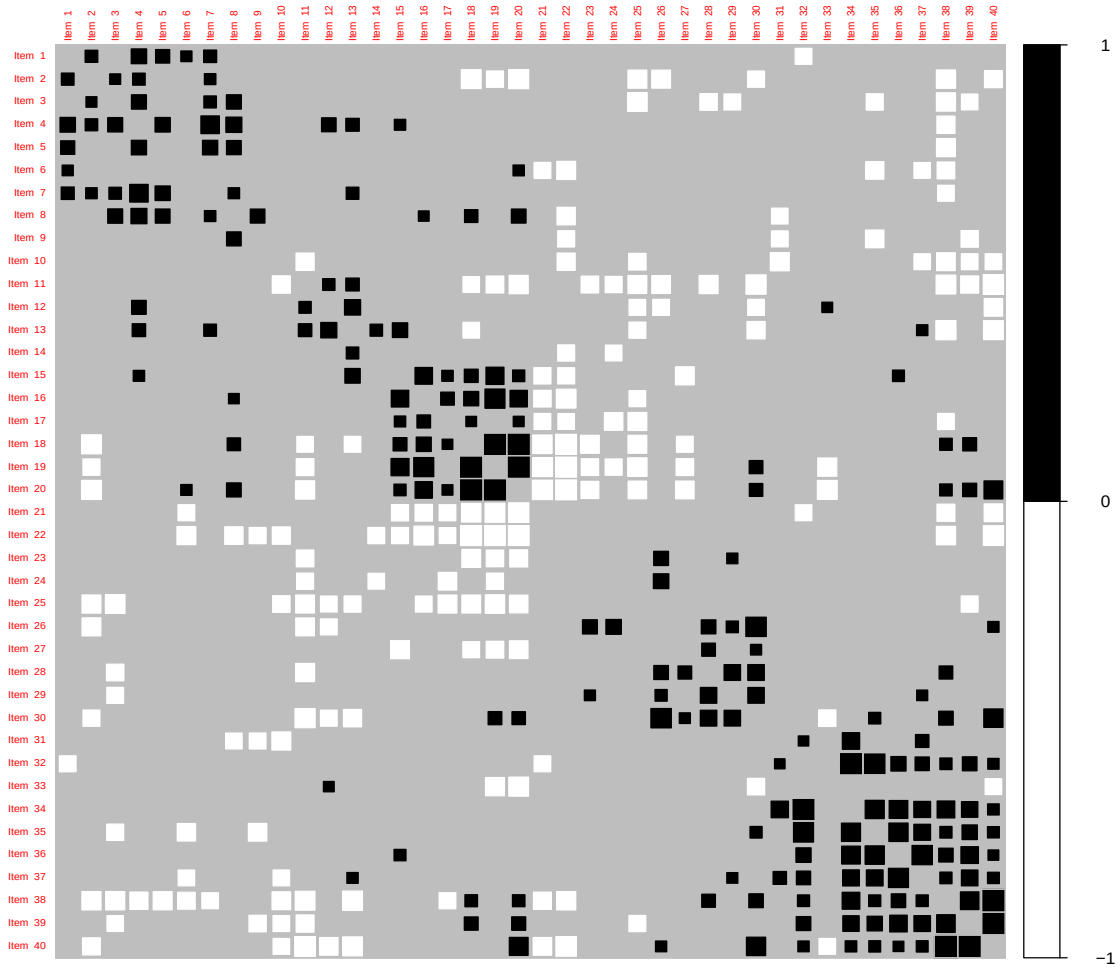

Figure 10: A Plot of the values of the Z_{LI} statistic for the Mathematics data.

Example 2: A Licensure Test

Two data sets from a licensure test were analyzed in several chapters of Cizek and Wollack (2017). We consider one of these data sets, which includes item scores and response times of 1,629 examinees on one test form with 170 operational items that are dichotomously scored. Sinharay and Johnson (2019) fitted a joint model that includes the two-parameter logistic IRT model and the LNMRT to this data set to detect item preknowledge. Figure 11 shows the total time in minutes versus the raw score on the test. There is a negative correlation (of -0.25) between the total time and the raw score on the test and, unlike for the mathematics data, the quickest examinees (say, those who took less than 100 minutes) obtained large raw scores. About 10 examinees seem to have spent considerably more time than the rest, but information on why that happened was

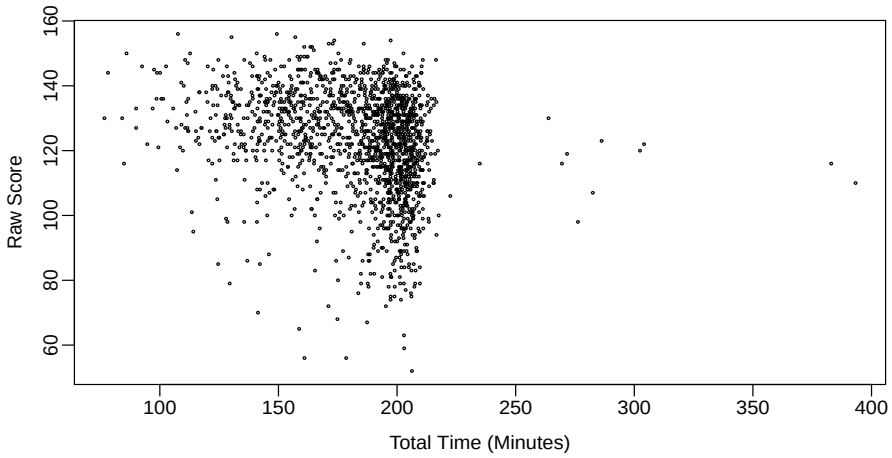

Figure 11: Plot of total time versus raw score for the Licensure data.

unavailable to the authors—accommodation is a possible explanation.[13]

The MLEs of the α_j's were between 1.37 and 2.70 and those of the β_j's were between 2.81 and 4.87, respectively. The Shapiro-Wilk test statistic (Shapiro & Wilk, 1965) and the Nikulin-Rao-Robson χ^2 statistic (Nikulin, 1973; Rao & Robson, 1974) were computed for all the items in the data set. The number of groups used was 30 for the latter test. At 5% level of significance, 73.5% and 67.1% of the items were found to have statistically significant values of the Shapiro-Wilk statistic and the Nikulin-Rao-Robson statistic, respectively. At 1% level, the percentages were 67.1% and 53.5%, respectively. The $S-\chi^2$ statistic (Orlando & Thissen, 2000) was significant for only 7.6% items for the data set and the statistic had a slightly negative relationship with the Nikulin-Rao-Robson statistic and the Shapiro-Wilk statistic. In addition, unlike the mathematics test, the correlation between the person-fit statistic of Sinharay (2018) and average response time is -0.02, which indicates that the model misfit in the data cannot be explained by quick responding. The Z_{LI} statistic was significant at 5% level for 30.9% item-pairs.

Figure 12 shows the normal probability plots of the log-response times for nine items (randomly chosen) among the 170 items. The figure shows some signs of departure

[13]A model-fit analysis after removing these examinees does not lead to much differences in the results—so these examinees are included in the analyses.

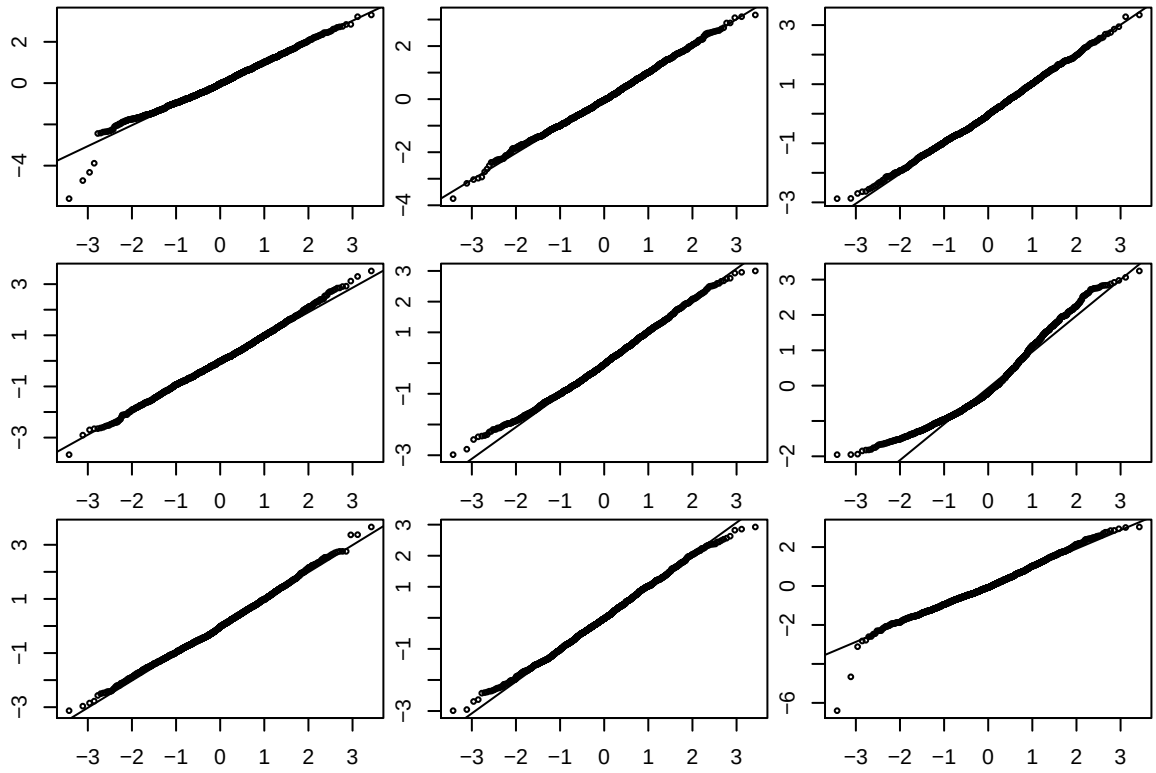

Figure 12: Normal probability plots for nine items from the Licensure data.

of the log-response times from a normal distribution, but the extent of departure is considerably less than that in Figure 7.

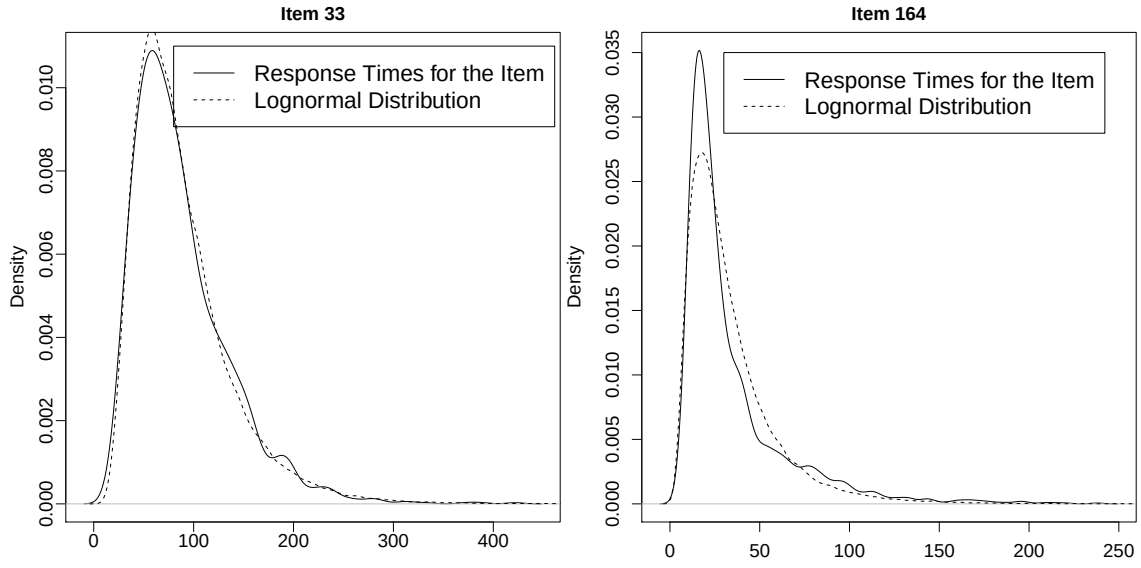

Figure 13: The Density of the Response Times for Two Items.

Figure 13 shows the density of the response times versus that of the best-fitting lognormal distribution for two items represented in Figure 12—Item 33 (Column 2 of the top row in Figure 12) and Item 164 (Column 3 of the middle row in Figure 12). These two items were picked because the model misfit appears not severe for the former and severe for the latter. While the density of the response times is close to that of the lognormal distribution for Item 33 (left panel), there is a substantial gap between the two curves for Item 164 (right panel). Several individuals take longer than what is expected from the LNMRT for the latter item[14] for which the mean and standard deviation of the response time are about 35 and 31, respectively.

Conclusions and Recommendations

This paper focuses on the LNMRT and suggests the use of several statistics for assessing item fit and local independence of the LNMRT. A simulation study demonstrates that the suggested statistics have satisfactory Type I error rate and power, especially when compared to the existing fit statistics. In general, the Shapiro-Wilk statistic and the Z_{LI} statistic had the largest power to detect item misfit and violation of local independence, respectively. Two real data applications demonstrate the usefulness of the statistics. Computer codes for computing the new fit statistics are also provided.

The item-fit statistics are based on the classical tests for normality (e.g., Thode, 2002) and the Nikulin-Rao-Robson (Nikulin, 1973) test. The test for local independence is based on standardized residuals in structural equation models. The asymptotic null distributions of all the suggested statistics are known and/or tables of critical values for the test statistics are publicly available. The simplicity of the suggested methodologies and their strong theoretical basis (in the form of asymptotic null distributions) promise to make them attractive to those interested in assessing the goodness of fit of response-time models.

The LNMRT was found to offer inadequate fit to two real data sets—one from a grade 8 mathematics test and one from a licensure test. The percentages of statistically significant values of the item-fit statistics and Z_{LI} were much larger than the nominal

[14]This is most clear for values of time about 100

level for both of these data sets. Similar poor fit was observed for two other real data sets from tests that have a time limit (results not discussed and can be obtained from the lead author). This result of poor fit of the LNMRT for multiple real data sets is important given the observation of Bolsinova and Tijmstra (2018, p. 13) that the LNMRT is used in most applications of response-time modeling.

Researchers such as Gelman et al. (2014, p. 151) noted that finding an extreme p-value and thus rejecting a model is never the end of an analysis. Therefore, a natural question in the context of this paper is "What should a practitioner do when a misfit of the LNMRT is found?" It is possible to do several things if a misfit is found. First, as Gelman et al. (2014, p. 151) stated, one can look for other models, including extensions of the current model, that may improve the fit. In our context, the Box-Cox normal model for the response times is a possible extension of the LNMRT that was found to fit response-times data better by Klein Entink et al. (2009); the extension of the LNMRT suggested by Bolsinova and Tijmstra (2018) is another possible candidate. Second, given that the data examples show the presence of some outlying/aberrant examinees, a simple extension may not fit the data and one may need to fit a mixture response-time model (that assumes one model for normal responses and another model for aberrant responses) such as that of Wang, Xu, Shang, and Kuncel (2018). Third, as noted by researchers such as Sinharay and Haberman (2014), practitioners should assess practical significance of any model misfit—such an assessment aims to answer questions such as "Are the main inferences made from the model influenced by the model misfit?" and "Can the model, with its misfit, still be used for the present problem?" The assessment of practical significance is problem-specific and depends heavily on the purpose for which the model is being used. An example of such an analysis would be that in an application of the LNMRT to person-fit analysis using response times (as in, for example, Sinharay, 2018), one finds 10% misfitting examinees, but then applies the Box-Cox normal model (Klein Entink et al., 2009) to find the percent of misfitting examinees; if only 5% examinees are found misfitting with the Box-Cox normal model, then the misfit of the LNMRT is statistically significant.

Our paper has several limitations. First, applications of the suggested statistics to

more simulated and real data would provide more insight into these statistics. Second, extension of the suggested statistics to more complicated response-time models is a possible topic for future research. Third, other types of item-fit statistics for these models may be helpful. For example, the suggested statistics have low power under certain conditions and research on finding more powerful statistics will be useful. Finally, research on finding effect sizes corresponding to the suggested statistics would be useful. One way to examine effect size would be to find out the practical significance of the misfit (Sinharay & Haberman, 2014).

References

Adefisoye, J. O., Golam Kibria, B. M., & George, F. (2016). Performances of several univariate tests of normality: An empirical study. *Journal of Biometrics & Biostatistics*, *7*(4), 1–8.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, *49*, 765–769.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 13–38.

Boughton, K., Smith, J., & Ren, H. (2017). Using response time data to detect compromised items and/or people. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 177–190). Washington, DC: Routledge.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York, NY: Wiley.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of detecting cheating on tests.* Washington, DC: Routledge.

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, *58*, 341.

D'Agostino, R. B. (1986). Tests for the normal distribution. In R. B. D'Agostino & M. A. Stephens (Eds.), *Goodness-of-fit techniques* (pp. 367–420). New York, NY: Marcel Dekker.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*, 1–11. (doi=10.3389/fpsyg.2019.00102)

Finger, M. S., & Chee, C. S. (2009, April). *Response-time model estimation via confirmatory factor analysis.* Paper presented at the Annual meeting of the National Council of Measurement in Education, San Diego, CA.

Fox, J.-P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, *54*, 243–262.

Gan, F. F., & Koehler, K. J. (1990). Goodness-of-fit tests based on p-p probability plots. *Technometrics*, *32*, 289.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). New York, NY: Chapman and Hall.

Glas, C. A. W., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, *63*, 603–626.

Gross, J., & Ligges, U. (2015). *nortest: Tests for normality.* (R package version 1.0-4)

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*, 255–259.

Joreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*, 443–482.

Klein Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48.

Klein Entink, R. H., van der Linden, W. J., & Fox, J. P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*, 621–640.

Kyllonen, P., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(14), 1–29.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359–379.

Lilliefors, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*, 399–402.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*, 519–530.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*, 426–451.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, *58*, 445–469.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633.

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*, 533–558.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*, 56–74.

Moore, D. S. (1986). Tests of chi-square type. In R. B. D'Agostino & M. A. Stephens (Eds.), *Goodness-of-fit techniques* (pp. 63–96). New York, NY: Marcel Dekker.

Nikulin, M. S. (1973). Chi-square test for continuous distributions with shift and scale parameters. *Theory of Probability & Its Applications*, *18*, 559–568.

Ogasawara, H. (2001). Standard errors of fit indices using residuals in structural equation

modeling. *Psychometrika*, *66*, 421–436.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157–175.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38–47.

Ranger, J., & Kuhn, J. (2014). Testing fit of latent trait models for responses and response times in tests. *Psychological Test and Assessment Modeling*, *56*, 382–404.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley.

Rao, K. C., & Robson, D. S. (1974). A chi-square statistic for goodness-of-fit within the exponential family. *Communications in Statistics*, *3*, 1139–1153.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Anlytics*, *2*, 21–33.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*, 131–151.

Scheffe, H. (1959). *The analysis of variance*. New York, NY: Wiley.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

Schnipke, D. L., & Scrams, D. J. (1999). *Representing response-time information in item*

banks (LSAC-R-97-09). Newtown, PA: Law School Admission Council.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Seber, G. A. F. (1984). *Multivariate observations*. New York, NY: John Wiley.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591.

Shapiro, S. S., Wilk, M. B., & Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, *63*, 1343–1372.

Singh, A. C. (1987). On the optimality and a generalization of Rao-Robson's statistic. *Communications in Statistics - Theory and Methods*, *16*, 3255–3273.

Sinharay, S. (2018). A new person-fit statistic for the lognormal model for response times. *Journal of Educational Measurement*, *55*, 457–476.

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*(1), 23–35.

Sinharay, S., & Johnson, M. S. (2019). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology*. (Advance online publication. doi:10.1111/bmsp.12187)

Skorupski, W. P., & Wainer, H. (2017). The case for Bayesian methods when investigating test fraud. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 214–231). Washington, DC: Routledge.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 179–203). New York, NY: Academic Press.

Thode, H. (2002). *Testing for normality*. New York, NY: Marcel Dekker.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272.

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*, 120–139.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347.

van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339–356.

van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *70*, 317–345.

Voinov, V., Nikulin, M., & Balakrishnan, N. (2013). *Chi-squared goodness of fit tests with applications*. New York, NY: Elsevier.

Voinov, V., Pya, N., & Alloyarova, R. (2009). A comparative study of some modified chi-squared tests. *Communications in Statistics - Simulation and Computation*, *38*, 355–367.

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*, 469–501.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wei, T., & Simko, V. (2017). *R package "corrplot": Visualization of a correlation matrix.* (Version 0.84)

Weisberg, S., & Bingham, C. (1975). An approximate analysis of variance test for non-normality suitable for machine calculation. *Technometrics*, *17*, 133.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183.

Wollack, J. A., & Schoenig, R. W. (2018). Cheating. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 260–265). Thousand Oaks, CA: Sage.

Yazici, B., & Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, *77*, 175–183.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psycholgoical Measurement*, *8*, 125–145.

Appendix A: R Functions to Compute the MLEs and the New Statistics

```
# Use of the R package 'lavaan' to fit the LNMRT
library(lavaan)
dy <- data.frame(y)# Data set (of dimension IxJ) with log-response times
J=10#J, the number items, is assumed to be 10 for the data set dy
model=paste("f1=~",paste0("a*X",1:(J-1),"+",collapse=""),paste("a*X",J,sep=""))
fit <- cfa(model, data = dy, meanstructure = TRUE, auto.var= TRUE)
pars <- coef(fit) #pars[12:21], sqrt(1/pars[1:10]), and pars[11]
                  #include the estimated beta's, alpha's, and sigma_squared
# Compute p-values for the Shapiro-Wilk test
ShapWilk=function(ltimes)
{nitem=ncol(ltimes)
 SW=rep(0,nitem)
 for (i in 1:nitem)
 {times=ltimes[,i]
  SW[i]=shapiro.test(times)$p.value}
  return(SW)}
# Compute p-values for the Anderson-Darling test
library(nortest)
AnDa=function(ltimes)#Anderson-Darling test
{nitem=ncol(ltimes)
 ad=rep(0,nitem)
 for (i in 1:nitem)
 {times=ltimes[,i]
  ad[i]=ad.test(times)$p.value}
  return(ad)}
# Compute the Nikulin-Rao-Robson Item-fit Statistic
NRR=function(ltimes,G)#G is the number of groups
{nitem=ncol(ltimes)
 n=nrow(ltimes)
 eo=EpsOmegas(G)
 eps=eo[,1]
 omega=eo[,2]
 chisq=rep(0,nitem)
 p=(1:(G-1))/G
 for (i in 1:nitem)
 {times=ltimes[,i]
  q=mean(times)+(sd(times))*qnorm(p)
  z=c(-100,q,100)
  obs=rep(0,G)
  for (j in 1:G)
  {obs[j]=length(times[times>z[j] & times<z[(j+1)]])
   chisq[i]=chisq[i]+(obs[j])**2}
```

```
   chisq[i]=G*chisq[i]/n-n+(sum(obs*eps))**2/n+(sum(obs*omega))**2/n }
  return(chisq)}
#A function that calculates some constants---these are input to the function NRR
 EpsOmegas=function(G)
 {y=c(-1000,qnorm((1:(G-1))/G),1000)
  p1=dnorm(y[1:G])
  p2=dnorm(y[2:(G+1)])
  a=p2-p1
  b=y[1:G]*p1-y[2:(G+1)]*p2
  lam1=1-G*sum(a*a)
  lam2=2-G*sum(b*b)
  return(cbind(G*a/sqrt(lam1),G*b/sqrt(lam2)))}
# Compute the Z_LI statistic for all item pairs
TestLI=function(ltimes)
{n=ncol(ltimes)
 colnames(ltimes)=paste("X",1:n,sep="")
 mod=paste("f1=~",paste0("a*X",1:(n-1),"+",collapse=""),paste("a*X",n,sep=""))
 fit <- cfa(mod, data = ltimes, meanstructure = TRUE, auto.var= TRUE)
 lr=lavResiduals(fit,type="cor.bollen")
 Zstat=lr$cov.z
 diag(Zstat)=0
 return(Zstat)} #Zstat: JxJ matrix consisting of the Z_LI for all item pairs
```

Appendix B: Type I Error Rates of the Statistics in the Simulation Study

Table A1: Type I Error Rates (as percentages) of the Shapiro-Wilk statistic (W), Anderson-Darling statistic (A^2), Nikulin-Rao-Robson χ^2 statistic (χ^2_N), Ranger-Kuhn's statistic (T_j), the Lagrange multiplier item-fit statistic (LM_j), the Lagrange multiplier local-independence statistic (LM_{LI}), and the Z_{LI} statistic.

Statistic	20 Items			40 Items			60 Items		
	500	1,000	5,000	500	1,000	5,000	500	1,000	5,000
W	4.9	5.1	4.5	5.0	5.3	4.2	5.0	5.3	4.0
A^2	5.2	4.5	4.5	5.2	5.3	4.7	5.1	5.1	4.0
χ^2_N	5.3	4.9	4.5	4.9	5.2	4.5	4.8	5.1	4.7
T_j	4.9	4.9	5.0	4.9	5.1	5.0	5.0	4.9	5.0
LM_j	6.5	6.6	6.5	6.3	6.1	6.6	6.5	6.2	6.9
LM_{LI}	6.5	6.3	6.6	6.3	6.5	6.6	6.6	6.8	6.9
Z_{LI}	4.7	4.7	4.5	4.8	5.0	5.0	5.0	4.9	4.9

Note: The three numbers—500, 1,000, and 5,000—at the top denote the sample sizes.

Table A1 shows the Type I error rates (at 5% level of significance) as percentages, rounded to the first decimal place, of the item-fit and local-independence statistics. The table shows that except for the LM_j and LM_{LI} statistics, all the other statistics have satisfactory Type I error rates—the rates are close to and often smaller than the nominal level in all simulation conditions. The conclusions on Type I error rates at 1% level (not reported here) are very similar.